Heterogenitätsbehandlung und Terminology Mapping durch Crosskonkordanzen – eine Fallstudie

Vivien Petras

GESIS-IZ Sozialwissenschaften, Bonn

This is an author's accepted manuscript version of a conference paper published in *Jörn Sieglerschmidt and H. Peter Ohly (Ed.), Wissensspeicher in digitalen Räumen: Nachhaltigkeit, Verfügbarkeit, semantische Interoperabilität. Proceedings der 11. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO), Konstanz, 20. bis 22. Februar 2008 (Vol. 11, pp. 91-102). Fortschritte in der Wissensorganisation. Würzburg: ERGON Verlag.*

Heterogenitätsbehandlung und Terminology Mapping durch Crosskonkordanzen – eine Fallstudie

Vivien Petras
GESIS-IZ Sozialwissenschaften, Bonn

Abstract

The German Federal Ministry for Education and Research (BMBF) funded a project until the end of 2007 with the objective of organizing the creation and management of cross-concordances between controlled vocabularies (thesauri, classifications, subject heading lists).

In three years, 64 cross-concordances with more than 500,000 relations between controlled vocabularies from the social sciences and other subject areas were established. In the final phase of the project, a major evaluation effort to test the effectiveness of the cross-concordances in different information systems was conducted.

The article reports on application areas of this heterogeneity treatment through cross-concordances and the results of the extensive analyses.

Zusammenfassung

Das BMBF hat bis Ende 2007 ein Projekt gefördert, deren Aufgabe es war, die Erstellung und das Management von Crosskonkordanzen zwischen kontrollierten Vokabularen (Thesauri, Klassifikationen, Deskriptorenlisten) zu organisieren.

In drei Jahren wurden 64 Crosskonkordanzen mit mehr als 500.000 Relationen zwischen kontrollierten Vokabularen aus den Sozialwissenschaften und anderen Fachgebieten umgesetzt. In der Schlussphase des Projekts wurde eine umfangreiche Evaluation durchgeführt, die die Effektivität der Crosskonkordanzen in unterschiedlichen Informationssystemen testen sollte.

Der Artikel berichtet über die Anwendungsmöglichkeiten der Heterogenitätsbehandlung durch Crosskonkordanzen und die Ergebnisse der umfangreichen Analysen.

1. Einführung: Heterogenität und das KoMoHe Projekt

Die Koordination der von der Deutschen Forschungsgemeinschaft (DFG) geförderten Virtuellen Fachbibliotheken und den vom Bundesministerium für Bildung und Forschung (BMBF) unterstützten Informationsverbünden führte zur Schaffung eines generellen Wissenschaftsportals (www.vascoda.de), in dem eine zentrale Suche zu Fachportalen und Fachclustern weiterleitet (Depping 2007). Diese Bündelung von Ressourcen schaffte hochkomplexe Strukturen und Anforderungen zur Integration der Informationsangebote - sowohl auf inhaltlicher als auch auf organisatorisch-technischer Ebene – die weit über unverbunden entwickelte Lösungsmodelle hinausgehen.

Vom September 2004 bis Ende 2007 wurde vom BMBF ein Kompetenzzentrum

Modellbildung und Heterogenitätsbehandlung (Projekt KoMoHe 2008) gefördert, daß die Entwicklung von vascoda maßgeblich unterstützen sollte. Das KoMoHe Projekt am GESIS-IZ Sozialwissenschaften in Bonn verfolgte zwei Ziele: (1) die übergreifende Modellbildung wissenschaftlicher Informationsangebote am konkreten Beispiel vascoda mit allen nachgeschalteten Ebenen (Krause / Mayr 2006), sowie (2) die Behandlung von Fragen zur Heterogenitätsbehandlung zur Förderung der semantischen Integration heterogener Informationssammlungen und als Ergänzung zur Standardisierung durch einheitliche Metadaten. Dieser Artikel beschreibt die Behandlung des zweiten Zieles (Heterogenitäts-behandlung).

Die Vereinheitlichung von formalen Metadaten für die verteilte Suche unterliegt historisch einer verbreiteten Standardisierung, u.a. durch die Anforderungen eines maschinellen Austausches von Bibliotheksdaten vorangetrieben. Das Dublin Core Metadatenformat (Dublin Core Metadata Element Set 2008) sowie die Austauschprotokolle OAI-PMH (Protocol for Metadata Harvesting 2008) und Z39.50 (National Information Standards Organization 2003) sind bekannte und breit angewendete Beispiele.

Auch die semantische Heterogenitätsbehandlung erfährt immer mehr Aufmerksamkeit von Anwendern und Forschern gleichermaßen (Chan / Zeng 2006, Doerr 2001, Zeng / Chan 2004, 2006). Für das zuverlässige und präzise Auffinden von Dokumenten aus verschiedenen Datenbanken müssen die Benennungen von gesuchten Konzepten in den verschiedenen Informationsressourcen bekannt sein. Sowohl Sucher als auch verschiedene Informationsressourcen (Datenbanken) benutzen unterschiedliche Terminologien, um Konzepte darzustellen. In einer verteilten Suche, wenn mehr als eine Datenbank abgesucht werden soll, kann es dazu führen, daß die unterschiedlichen Sprachen (Suchsprache, Dokumentationssprachen) ein Match verhindern, so daß relevante Dokumente aufgrund von unterschiedlichen Bezeichnungen nicht gefunden werden. Dies ist ein besonderes Problem in Datenbanken mit textlich spärlichen Datensätzen wie zum Beispiel Bibliothekskataloge oder bibliographische Datenbanken, die zur Inhaltsbeschreibung meist nur Schlagworte oder Zusammenfassungen (Abstracts) zur Verfügung haben. Selbst Volltextdatenbanken sind von diesem Problem der Sprachambiguität allerdings nicht gefeit. Unterschiedliche Benennungen tauchen selbstverständlich auch in mehrsprachigen Umgebungen auf, wo Dokumente und Suchanfragen aufgrund der unterschiedlichen benutzten natürlichen Sprachen nicht übereinstimmen. Ein Lösungsansatz für dieses Problem ist das Mapping von kontrollierten Vokabularen (Terminology Mapping), wie sie im KoMoHe Projekt unternommen wurden.

In den letzten Jahren haben unterschiedliche Institutionen Bemühungen im Bereich der semantischen Integration von Informationssystemen unternommen. In den Vereinigten Staaten hatte OCLC vor einigen Jahren das Terminology Services Project (2008) gegründet (Vizine-Goetz 2004, 2006) um Web Services für Terminologie Mappings zwischen unterschiedlichen kontrollierten Vokabularen wie z.B. DDC, LCC, LCSH oder MeSH zu testen und anzubieten. In Europa hat das Delos2 Network of Excellence in Digital Libraries Programm das Arbeitspaket 5 dem Problem von „Knowledge Extraction and Semantic Operability“ gewidmet (Patel et al. 2005). Im MACS Projekt wurden die drei großen nationalen Schlagwortsysteme LCSH (Englisch), Rameau (Französisch) und SWD (Deutsch) miteinander verbunden (Clavel-Merrin 2004). Ein anderes Projekt ist das CRISSCROSS Projekt (2008) an der Deutschen Nationalbibliothek und der Fachhochschule Köln, dass ein

multilinguales Mapping zwischen der Schlagwortnormdatei (SWD) und den Notationen der Dewey Dezimal Klassifikation (DDC) erstellt (Panzer 2008). Die High-Level Thesaurus Projekte (HILT 2008) an der University of Strathclyde sind ein weiteres Beispiel für die langjährige Entwicklung von Terminologie Mapping Technologien (Macgregor et al. 2007).

Die Frage der semantischen Interoperabilität wird weiterhin als Herausforderung im europäischen (siehe z.B. Arbeitspaket 2 des Europeana Projektes European Digital Library Network, daß sich mit „technical and semantic interoperability" auseinandersetzt) und internationalen Raum betrachtet. Im SKOS Standard (Simple Knowledge Organization System 2008) in Section 10 wird über Mapping Properties diskutiert, um auch im Semantic Web Mappingstrukturen hinreichend abbilden zu können.

2. *Implementation: Crosskonkordanzen am GESIS-IZ Sozialwissenschaften*

Die GESIS hat über mehrere Projekte unterschiedliche Terminology Mapping Verfahren verfolgt: statistische Verfahren (Hellweg et al. 2001) und intellektuelle Verfahren (Mayr / Walther 2008), deren Analyse der Fokus dieser Arbeit ist. Im KoMoHe Projekt wurde als Antwort auf die Integration heterogener Informationsressourcen die Erstellung von Crosskonkordanzen zwischen kontrollierten Vokabularen in den Mittelpunkt gestellt.

Crosskonkordanzen sind intellektuell erstellte, gerichtete, relevanz-bewertete Relationen zwischen den Termen zweier Wissensorganisationssysteme. Mit Wissensorganisations-systemen oder KOS (Knowledge Organization Systems) bezeichnen wir die Gesamtheit von kontrollierten Vokabularen, zwischen denen Relationen hergestellt werden können, also Thesauri, Klassifikationen, Schlagwortlisten, Taxonomien etc.

Wie in Abbildung 1 dargestellt, konnte das KoMoHe Projekt nicht alle Vokabulare gleichermaßen abbilden. Ein kleineres Vokabular, daß eine bestimmte Menge von Konzepten enthält, kann nur mit einer bestimmten Menge des Zielvokabulars relationiert werden, so daß in einem unilateralen Mapping bestimmte Termgruppen nicht berührt werden. Im Gegensatz dazu kann bei einem Mapping von einem größeren Vokabular zu einem kleineren Vokabular das Ausgangsvokabular nicht komplett zu Termen des Zielvokabulars gemappt werden.

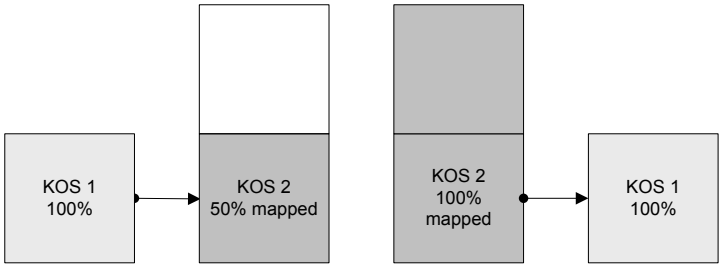

Abbildung 1: Unilaterale Mappings zwischen 2 kontrollierten Vokabularen (KOS) - Überlappung.

Zum Teil wurden auch nur Ausschnitte von Vokabularen in Crosskonkordanzen mit anderen Vokabularen verbunden. Aufgrund der Größenunterschiede war es notwendig,

bilaterale Crosskonkordanzen intellektuell von jeder Richtung (Ausgangsvokabular und Zielvokabular jeweils vertauscht) zu überprüfen.

Abbildung 2 zeigt einen weiteren Grund, warum die meisten Crosskonkordanzen nicht symmetrisch bilateral (d.h. von einer Richtung ausgehend und die Gegenrichtung automatisch gesetzt) erstellt wurden. In seltenen Fällen kann es vorkommen, daß aufgrund von unterschiedlichen Bedeutungsüberschneidungen von Konzepten in Vokabularen diese abhängig von der Konkordanzrichtung unterschiedlich gemappt werden.

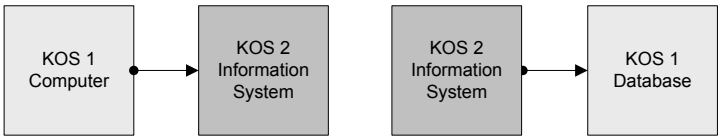

Abbildung 2: Unilaterale Mappings zwischen 2 kontrollierten Vokabularen (KOS) - Termunterschiede.

Viele unterschiedliche Relationen zwischen Konzepten der Vokabulare können repräsentiert werden. Abhängig vom Ziel der Crosskonkordanzen und den vorhandenen Ressourcen können die vielfältigsten semantischen Eigenschaften von Termrelationen unterschieden werden. Chaplan (1995) führt zum Beispiel 19 verschiedene Relationsarten in ihrem Mapping zwischen LCSH und dem Laborline Thesaurus auf. Im KoMoHe Projekt wurden 5 Relationsarten (siehe Tabelle 1) unterschieden: die Äquivalenzrelation, die Unterbegriffsrelation, die Oberbegriffsrelation, die Assoziationsrelation und die Null-Relation (keine Relation konnte festgestellt werden).

Vokabular A	**Relation**		**Vokabular B**
hacker	=	Äquivalenz	Hacking
hacker	^+	Assoziation	computers + crime
isdn device	0	Null	
isdn	<	Unterbegriff	Telecommunications
documentation system	>	Oberbegriff	abstracting services

Tabelle 1: Crosskonkordanz-Relationen Beispiele.

Von 2004 bis 2007 wurden im KoMoHe Projekt 16 Thesauri, 6 Schlagwortlisten und 3 Klassifikation in unterschiedlichen Kombinationen miteinander relationiert. Pro Vokabular wurden zwischen 1.000 und 17.000 Terme verbunden, wobei einige Vokabulare nur teilweise relationiert wurden. Die kontrollierten Vokabulare waren hauptsächlich deutschsprachige Terminologien, es wurden aber auch acht englische, ein russisches (INION) und drei mehrsprachige Vokabulare (Agrovoc, ELSST, Euro-Thesaurus) in Crosskonkordanzen verknüpft.

Viele der Crosskonkordanzen enthielten Terminologien aus den Sozialwissenschaften, da die GESIS eine Serviceeinrichtung für die Sozialwissenschaften ist. Allerdings wurden auch generelle Vokabulare sowie disziplinarisch angrenzende Vokabulare in dieses semantische Netz einbezogen. Abbildung 3 zeigt eine Übersicht dieses Crosskonkordanzennetzwerks.

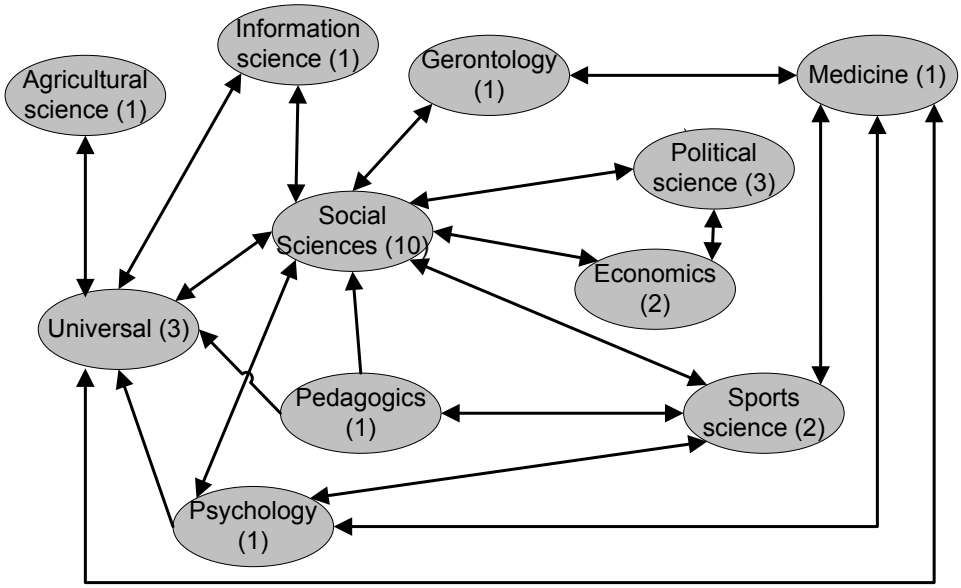

Abbildung 3: Netz der Terminologie-Mappings im KoMoHe Projekt. Die Zahlen und den Klammern enthalten die Anzahl der abgebildeten kontrollierten Vokabulare einer Disziplin.

3. Evaluation der Crosskonkordanzen

Crosskonkordanzen können den vielfältigsten Zwecken dienen. Eine Reihe von Fragen kann durch das Abfragen von individuellen Termtransformationen sowie die Analyse der Gesamtzahl aller Transformationen einer Crosskonkordanz beantwortet werden:

- Wieviele Ausdrücke können für ein Konzept gefunden werden?
- Welche Konzepte sind verwandt?
- Wie sieht das semantische Netzwerk einer Terminologie aus?
- Sind Vokabulare eher breiter oder enger in ihrer Abdeckung?
- Welche kontrollierten Vokabulare sind sich sehr ähnlich?
- Welche Disziplinen / Fachgebiete liegen beieinander und welche sind weiter voneinander entfernt?
- Welche Teile der Terminologien müssen relationiert, d.h. verbunden werden?

Einige Indizien gibt die quantitative Analyse der Crosskonkordanzen, aber die eigentliche Frage: „Welche Mappings sind am nützlichsten für die Suche?“ kann nur durch einen Retrieval-Test beantwortet werden. Im folgenden werden kurz die wichtigsten Kennziffern der KoMoHe Crosskonkordanzen beschrieben bevor der Aufbau und die Ergebnisse des am GESIS-IZ durchgeführten Information Retrieval Tests der Crosskonkordanzen erläutert werden.

3.1 Quantitative Analyse

Im KoMoHe Projekt wurden 25 kontrollierte Vokabulare in 28 bilateralen und 4 unilateralen (insgesamt 60) Crosskonkordanzen erstellt. Dabei wurden über eine halbe Million Relation zwischen Termen hergestellt. Tabelle 2 gibt einen kurzen Überblick über die Anzahl der Relationen und Terme.

Relationen	513.000
Äquivalenzrelationen	205.000
Ausgangsterme	380.000
Zielterme	200.000

Tabelle 2: Anzahl der Relationen, Äquivalenzrelationen und verbundene Ausgangs- und Zielterme aller Crosskonkordanzen.

Pro Crosskonkordanz wurden 6.500 Ausgangsterme zu 3.600 Zieltermen relationiert (1,2 Relationen pro Term) und 7.700 Relationen (930 Kombinationen) insgesamt erstellt. Durchschnittlich wurden 61% der Terme des Zielvokabulars verbunden. Abbildung 4 stellt die durchschnittliche Verteilung der Relationstypen dar, wobei diese von Crosskonkordanz zu Crosskonkordanz stark schwanken kann.

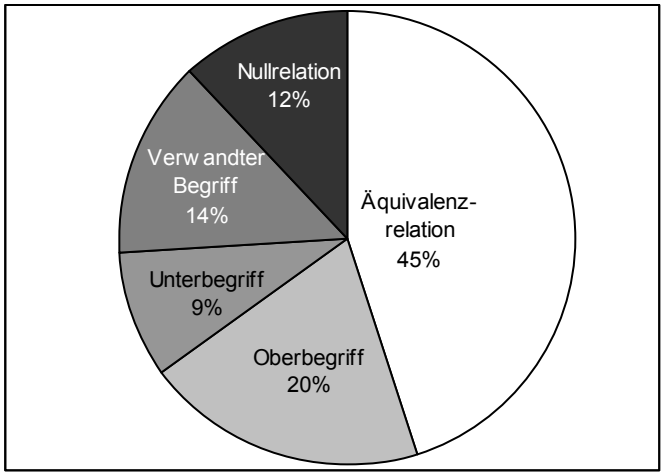

Abbildung 4: Verteilung der Relationstypen über alle Crosskonkordanzen.

Die Verteilung und Effektivität von Crosskonkordanzen kann sich nach Art der Crosskonkordanz sehr stark unterscheiden. Unterschiedlichen Einfluß auf die Suche haben die folgenden Einflußfaktoren:

- Disziplinwechsel (innerdisziplinär, interdisziplinär)
- Sprachenwechsel (einsprachig, zweisprachig)

- Vokabulargröße (gleich, Ausgangsvokabular > Zielvokabular, Ausgangsvokabular < Zielvokabular)
- Vergabehäufigkeit von Termen des Vokabulars in der Datenbank (entspricht der Popularität oder Geläufigkeit von Termen)

Ein Beispiel für die Wirkung von unterschiedlichen Crosskonkordanzen ist die Verteilung der Äquivalenzrelation. Die Äquivalenzrelation („dieser Term in Vokabular A" entspricht ungefähr diesem Term in Vokabular B") umfaßt 45% aller Relationen. Sie wird als wichtigste Relation gewertet, da sie direkt für die Übersetzung von Anfragetermen sowie für die automatische Anfrageerweiterung eingesetzt werden kann.

Es gibt drei unterschiedliche Typen von Äquivalenzrelationen: (1) ein identischer Term im Start- und Zielvokabular, (2) ein direktes Synonym, oder (3) eine Kombination von Termen im Zielvokabular (synonyme Kombination).

Identische Äquivalenzrelationen können keine Auswirkungen auf die Suche haben, da der Suchterm im Start- und Zielvokabular der gleiche ist, d.h. durch die Termtransformation wird keine neue Information erzeugt. Fast die Hälfte der Äquivalenzrelationen (46%) sind identische Relationen. Diese Verteilung schwankt allerdings stark, abhängig davon ob Ausgangs- und Zielvokabular der gleichen Disziplin entstammen oder interdisziplinäre Verbindungen darstellen. Die stärkere Präsenz von identischen Termtransformationen in innerdisziplinären Crosskonkordanzen läßt den Schluß zu, daß interdisziplinäre Crosskonkordanzen wahrscheinlich einen größeren Effekt für die Suche haben werden.

3.2 Information Retrieval Tests

Das Ziel der Information Retrieval Tests war es, den Effekt der Termtransformationen durch die Crosskonkordanzen auf die Recherche empirisch zu analysieren.

Grundsätzlich nehmen wir an, daß die Übersetzung der Anfrageterme in andere Vokabulare durch die Crosskonkordanzen die verteilte Suche in mehreren Datenbanken (mit unterschiedlichen Terminologien) ermöglicht. Durch die Anwendung der korrekten Terminologie und die damit präzisere Suche sollte sich eine Verbesserung des Recalls (Anzahl an relevanten Dokumenten) und womöglich auch der Precision (Genauigkeit der Resultsmenge) beobachten lassen. Außerdem könnte durch die verbesserte verteilte Suche auch eine Erhöhung der Dokumentendiversität eintreten, d.h. unterschiedliche Dokumenttypen können durch die Anwendung der verschiedenen Datenbankenvokabulare gefunden werden. Nichtsdestotrotz steht über diesen zu evaluierenden Effekten der Crosskonkordanzen der Versuch, das Sucherlebnis insgesamt ohne weiteren Aufwand für den Sucher zu verbessern.

In einem ersten Evaluationsschritt wurde die Effektivität der Crosskonkordanzen basierend auf den Maßzahlen Recall und Precision getestet. Anhand von realen Nutzeranfragen wurde der Mehrwert der Termtransformationen in der Anfrageerweiterung in der Suche untersucht. Ausgehend von der Suche in einer Datenbank wurde analysiert, ob mit Hilfe der Termtransformationen die Suche in einer zweiten Datenbank zu besseren Ergebnissen führt.

Dazu wurde zunächst eine Schlagwortsuche mit den Schlagwörtern der Ausgangsdatenbank in der Zieldatenbank mit einer Schlagwortsuche mit transformierten Schlagwörtern in der Zieldatenbank verglichen. Bleibt das Suchergebnis gleich, haben

Crosskonkordanzen und die Übersetzung von Suchtermen durch Termtransformationen keinen Effekt. Wird das Suchergebnis schlechter, hat der Einsatz der Crosskonkordanzen einen negativen Effekt. Wird das Suchergebnis besser, was unterstellt wird, haben die Crosskonkordanzen einen positiven Effekt in der Termerweiterung.

Test 1: Verbessert der Einsatz der Crosskonkordanzen die Schlagwortsuche?
CT (Ausgangsvok.) → Zieldatenbank
CT (Ausgangsvok.) → TT (Zielvok.) → Zieldatenbank
CT = Controlled Term TT = Termtransformation

In einem zweiten Test wurde auch der Effekt der Crosskonkordanzen für die Suche im Freitext (Überall-Suche) untersucht. Da die Crosskonkordanzen Schlagwörter oder Deskriptoren übersetzen, könnte man zunächst postulieren, dass die Crosskonkordanzen einen kleineren Effekt auf die Freitextsuche haben könnten, da nicht nur Schlagworte abgesucht werden.

Test 2: Verbessert der Einsatz von Crosskonkordanzen auch eine Freitextsuche?
FT →Zieldatenbank
FT → FT + TT (Zielvok.) → Zieldatenbank
FT = Free-text Term

Für alle Suchen wurden reale Nutzeranfragen, die von den Betreibern der untersuchten Datenbanken eingeholt wurden, benutzt. Für die Freitextsuche wurden ca. 1-3 natürlichsprachige Terme verwendet. Die Schlagwortsuche wurde mit von Informationsvermittlern operationalisierten Schlagwortanfragen (Boolsche Anfragen mit ca. 2-6 Termen) in den jeweiligen Datenbanken durchgeführt. Die Suchresultate wurden auf die Relevanz für eine Anfrage bewertet. Pro untersuchter Crosskonkordanz wurden zwischen 3-10 Nutzeranfragen getestet (durchschnittlich 6-7). Es wurden nur Äquivalenzrelationen für die Übersetzung der Terme benutzt. Abbildung 5 zeigt Beispiele für die Anwendung der Crosskonkordanzen in Tests 1 und 2.

Test 1: Verbessert der Einsatz der Crosskonkordanzen die Schlagwortsuche?
CT: Familienbeziehungen [*Ausgangsschlagwort*]
TT: Familien UND soziale Beziehungen [*Termtransformation durch Crosskonkordanz*]

Test 2: Verbessert der Einsatz von Crosskonkordanzen auch eine Freitextsuche?
FT: Familienbeziehungen (*Ausgangsfreitextsuche*)
FT+TT: Familienbeziehungen ODER (Familien UND soziale Beziehungen) [*Ausgangsfreitextsuche ODER Termtransformation durch Crosskonkordanz*]

Abbildung 5. Beispiele für Information Retrieval Tests 1 und 2.

Zur Evaluation des Effektes der Crosskonkordanzen auf die Suche wurden klassische Information Retrieval-Messwerte betrachtet:

- Recall: Anteil gefundener relevanter Dokumente von allen relevanten Dokumenten (Durchschnitt über alle Anfragen pro Suchvariante)
- Precision: Anteil gefundener relevanter Dokumente von allen gefundenen Dokumenten (Durchschnitt über alle Anfragen pro Suchvariante)
- P10: Precision at 10 = Precision nach 10 gefundenen Dokumenten
- P20: Precision at 20 = Precision nach 20 gefundenen Dokumenten

In den Tabellen 3 und 4 sieht man einen Überblick über die prozentuale Veränderung der Maßzahlen wenn Crosskonkordanzen für die Suche eingesetzt werden. Die Zahlen in Klammern in der linken Spalte enthalten die Anzahl der untersuchten Crosskonkordanzen.

Das Resultat des Einsatzes der Crosskonkordanzen in der Schlagwortsuche (Tabelle 3) ist überwältigend positiv. In der Suchvariante mit Termtransformationen wird die Anzahl der gefundenen Dokumente mehr als verdoppelt. Recall, also die Menge der gefundenen relevanten Dokumente, erhöht sich fast um 100% und Precision, also die Menge der gefundenen Dokumente, die relevant sind, erhöht sich um fast 50%.

Anhand der Messwerte kann man auch den Unterschied zwischen innerdisziplinären und interdisziplinären Crosskonkordanzen feststellen. Eine außergewöhnliche Verbesserung kann man bei den Crosskonkordanzen zwischen verschiedenen Disziplinen beobachten. Hier erhöhen sich der Recall und die Precision um mehr als den Durchschnitt.

	Recall	Precision	P@10	P@20
Alle (13)	+91,8%	+53,29%	+ 53,6%	+62,9%
Innerdisziplinär (5)	+39,3%	+33,9%	+39,1%	+43,7%
Interdisziplinär (8)	+135,6%	+67,8%	+62,8%	+73,7%

Tabelle 3: Ergebnisse der Test 1 Evaluation. Veränderungen der Maßzahlen bei Verwendung von TT gegenüber CT in Prozent.

Wie man in Tabelle 4 erkennen kann, erhöht sich der Recall bei Einsatz der Crosskonkordanzen auch in der Freitextsuche. Allerdings verringert sich die Precision, d.h. das Ergebnis enthält auch mehr nicht relevante Dokumente. Für die innerdisziplinären Crosskonkordanzen verringert sich die Precision weniger als für die interdisziplinären Crosskonkordanzen. Das mag z. T. damit zusammenhängen, dass zu wenige interdisziplinäre Crosskonkordanzen (2) untersucht wurden und man keinen signifikanten Trend erkennen kann.

	Recall	Precision	P@10	P@20
Alle (8)	+20,7%	-13,6%	-21,9%	-16,3%
Innerdisziplinär (6)	+19,6%	-11,5%	-17,5%	-10,0%
Interdisziplinär (2)	+23,8%	-23,7%	-46,7%	-47,5%

Tabelle 4: Ergebnisse der Test 2 Evaluation. Veränderungen der Maßzahlen bei Ansatz von FT+TT gegenüber FT in Prozent.

Die positiven Effekte des Einsatzes der Crosskonkordanzen für die Suche in heterogen erschlossenen Datenbanken konnte im oben beschriebenen Testszenario sehr deutlich empirisch bestätigt werden. Sowohl für inner- als auch interdisziplinäre Crosskonkordanzen konnte eine erhebliche Verbesserung der Suchsituation nachgewiesen werden. Die

Ergebnismenge ist nicht nur größer, sondern kann auch präziser werden (gemessen in Recall und Precision).

4. Ausblick: Anwendungen in digitalen Bibliotheken

Die Crosskonkordanzen der KoMoHe und CARMEN Projekte (IZ Sozialwissenschaften 2002) wurden in eine relationale Datenbank eingepflegt und können dort problemlos abgerufen und aktualisiert werden. Zudem wurde ein Web Service (Heterogenitätsservice) entwickelt, der in der jetzigen Konfiguration automatisch alle äquivalenten Termtransformationen für einen gegebenen Ausgangsterm ausgibt.

Der Heterogenitätsservice wird eingesetzt für die automatische Anfrageerweiterung im sozialwissenschaftlichen Fachportal Sowiport (http://www.sowiport.de) und befindet sich auch im Testbetrieb für das allgemeine Wissenschaftsportal vascoda.

Die Crosskonkordanzen werden auch im multilingualen Retrieval und zur Termerweiterung in der Information Retrieval Evaluationsinitiative CLEF (Cross Language Evaluation Forum, s. Petras et al. 2007) benutzt.

Als nächste Schritte streben wir eine verbesserte Visualisierung der Transformationsergebnisse der Crosskonkordanzen an sowie die Konvertierung der Crosskonkordanzen Daten in das SKOS Format, daß in diesem Jahr vom W3C standardisiert werden soll und dann als generelles Austauschformat für kontrollierte Vokabulare und Terminology Mappings dienen kann.

Das Potential der Crosskonkordanzen für die Recherche ist noch längst nicht ausgeschöpft. Mit nur 23% aller kreierten Termtransformationen (nicht-identische Äquivalenzrelationen) konnten wir erstaunliche Ergebnisverbesserungen erzielen. Die Effektivität der anderen Relationen wurde dabei noch nicht betrachtet. Weitere wichtige Perspektiven können die semantischen Beziehungen des gesamten Crosskonkordanzen-Netzwerks liefern, das auch eine Darstellung der verschiedenen Wissenschaftssprachen repräsentiert. Die Beziehungen der Vokabularien untereinander und die indirekten Verbindungen, die durch direkte Crosskonkordanzen entstehen können, bedürfen einer tieferen Analyse.

Literatur

Chan, L. M., & Zeng, M. L. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part I: Achieving Interoperability at the Schema Level. D-Lib Magazine, 12(6).

Chaplan, M. A. (1995). Mapping Laborline-Thesaurus Terms To Library-Of-Congress Subject-Headings - Implications For Vocabulary Switching. Library Quarterly 65(1): 39-61.

Clavel-Merrin, G. (2004). MACS (Multilingual Access to Subjects): a virtual authority file across languages. Cataloging & Classification Quarterly 39(1/2): 323-330

CRISSCROSS Projekt (2008). http://www.fbi.fh-koeln.de/institut/projekte/CrissCross/index_en.html (Last checked: 22-10-2008)

Depping, R. (2007). vascoda.de and the system of the German virtual subject libraries. In A. R. D. Prasad & D. P. Madalli (Eds.), International Conference on Semantic Web & Digital Libraries (ICSD 2007) (pp. 304-314). Bangalore, India: Documentation Research & Training Centre, Indian Statistical Institute.

Doerr, M. (2001). Semantic Problems of Thesaurus Mapping. Journal of Digital Information, 1(8).

Dublin Core Metadata Element Set (2008). http://dublincore.org/documents/dces/ (Last checked: 22-10-

2008)
Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M. N. O., Mutschke, P., et al. (2001). Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften.
HILT (2008). http://hilt.cdlr.strath.ac.uk/ (Last checked: 22-10-2008)
IZ Sozialwissenschaften (2002). CARMEN: Content Analysis, Retrieval and MetaData: Effective Networking. Abschlussbericht des Arbeitspakets 12 (AP 12) Crosskonkordanzen von Klassifikationen und Thesauri**:** 51.
Krause, Jürgen; Mayr, Philipp (2006). Allgemeiner Bibliothekszugang und Varianten der Suchtypologie - Konsequenzen für die Modellbildung in vascoda. Bonn: Informationszentrum Sozialwissenschaften. 52 p. (IZ-Arbeitsbericht Nr. 38) http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_38.pdf (Last checked: 22-10-2008)
Macgregor, G., Joseph, A., & Nicholson, D. (2007). A SKOS Core approach to implementing an M2M terminology mapping server. Bangalore, India.
Mayr, P., & Walter, A.-K. (2008). Mapping Knowledge Organization Systems. In H. P. Ohly, S. Netscher & K. Mitgutsch (Eds.), Fortschritte der Wissenorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation (pp. 80-95). Würzburg: Ergon.
National Information Standards Organization (U.S.) (2003). Information retrieval (Z39.50) : application service definition and protocol specification : an American national standard. National information standards series. http://www.loc.gov/z3950/agency/Z39-50-2003.pdf (Last checked: 22-10-2008)
Panzer, M. (2008). Semantische Integration heterogener und unterschiedlichsprachiger Wissensorganisationssysteme: CrissCross und jenseits. In H. P. Ohly, S. Netscher & K. Mitgutsch (Eds.), Fortschritte in der Wissensorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation (pp. 61-69). Würzburg: Ergon.
Patel, M., Koch, T., Doerr, M., & Tsinaraki, C. (2005). Semantic Interoperability in Digital Library Systems.
Petras, V., Baerisch, S., & Stempfhuber, M. (2007). The Domain-Specific Track at CLEF 2007, Cross Language Evaluation Forum Workshop (CLEF) 2007. Budapest.
Projekt KoMoHe (2008). http://www.gesis.org/forschung-lehre/programme-projekte/informationswissenschaften/projektuebersicht/komohe/ (Last checked: 22-10-2008)
Protocol for Metadata Harvesting (2008). http://www.openarchives.org/pmh/ (Last checked: 22-10-2008)
Terminology Services Project (2008). http://www.oclc.org/research/projects/termservices/ (Last checked: 22-10-2008)
Simple Knowledge Organization System (2008). http://www.w3.org/TR/2008/WD-skos-reference-20080829/#mapping (Last checked: 22-10-2008)
Vizine-Goetz, D., Hickey, C., Houghton, A., & Thompsen, R. (2004). Vocabulary Mapping for Terminology Services. Journal of Digital Information, 4(4).
Vizine-Goetz, D., Houghton, A., & Childress, E. (2006). Web Services for Controlled Vocabularies. ASIS&T Bulletin, 2006 (June/July)
Zeng, M. L., & Chan, L. M. (2004). Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. Journal of the American Society for Information Science and Technology, 55(3), 377-395.
Zeng, M. L., & Chan, L. M. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels. D-Lib Magazine, 12(6).